globus

Globus: The Platform for Research Data Management

Vas Vasiliadis

ATPESC – August 6, 2021

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

Warning:
We may hear from unregistered attendees!

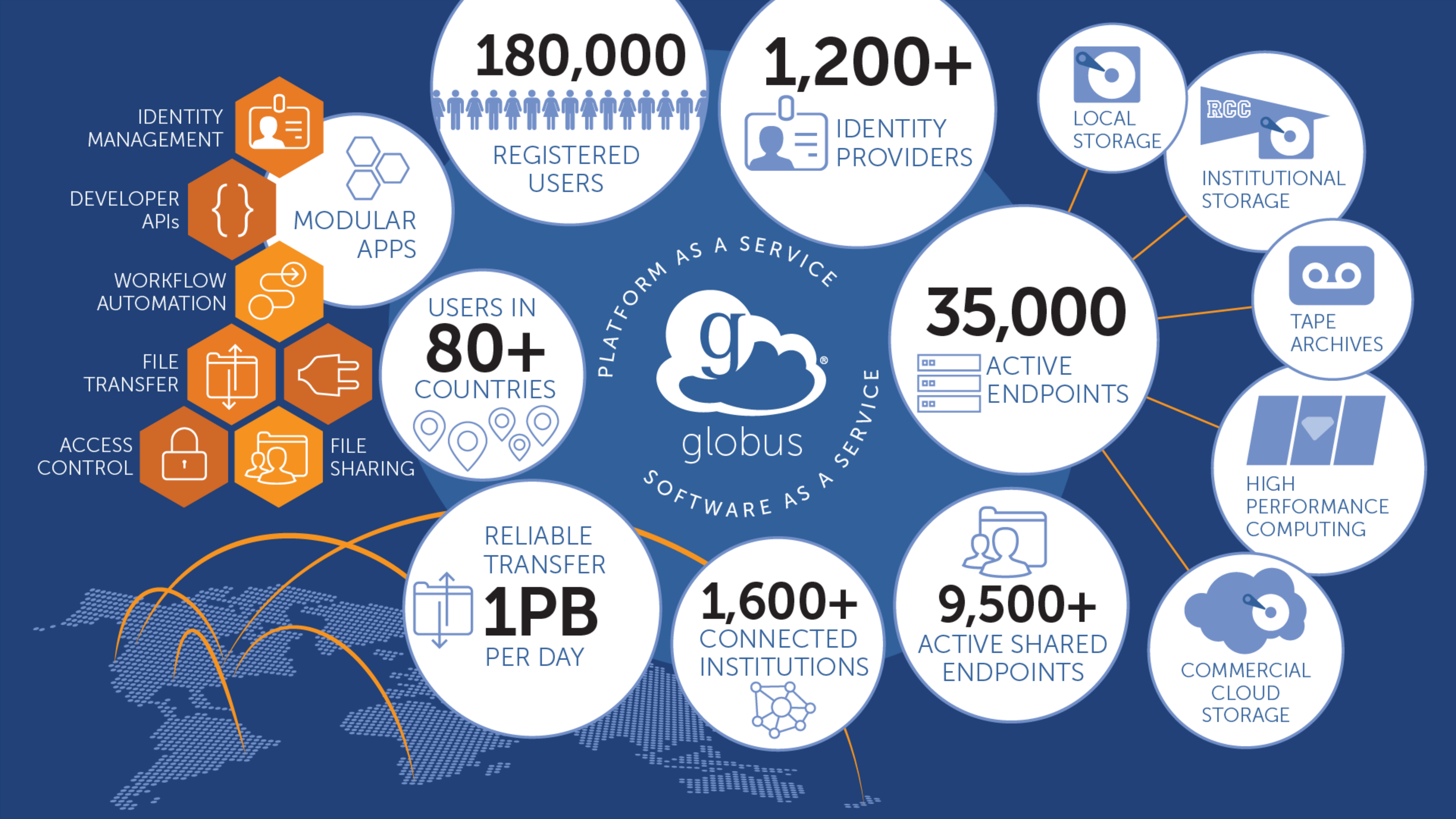
Globus is …

a non-profit service developed and operated by

THE UNIVERSITY OF CHICAGO

Our mission is to…

increase the efficiency and effectiveness of researchers engaged in data-driven science and scholarship through **sustainable** software



IDENTITY MANAGEMENT

DEVELOPER APIs

WORKFLOW AUTOMATION

FILE TRANSFER

ACCESS CONTROL

FILE SHARING

MODULAR APPS

180,000 REGISTERED USERS

1,200+ IDENTITY PROVIDERS

USERS IN 80+ COUNTRIES

PLATFORM AS A SERVICE

globus®

SOFTWARE AS A SERVICE

35,000 ACTIVE ENDPOINTS

LOCAL STORAGE

INSTITUTIONAL STORAGE

TAPE ARCHIVES

HIGH PERFORMANCE COMPUTING

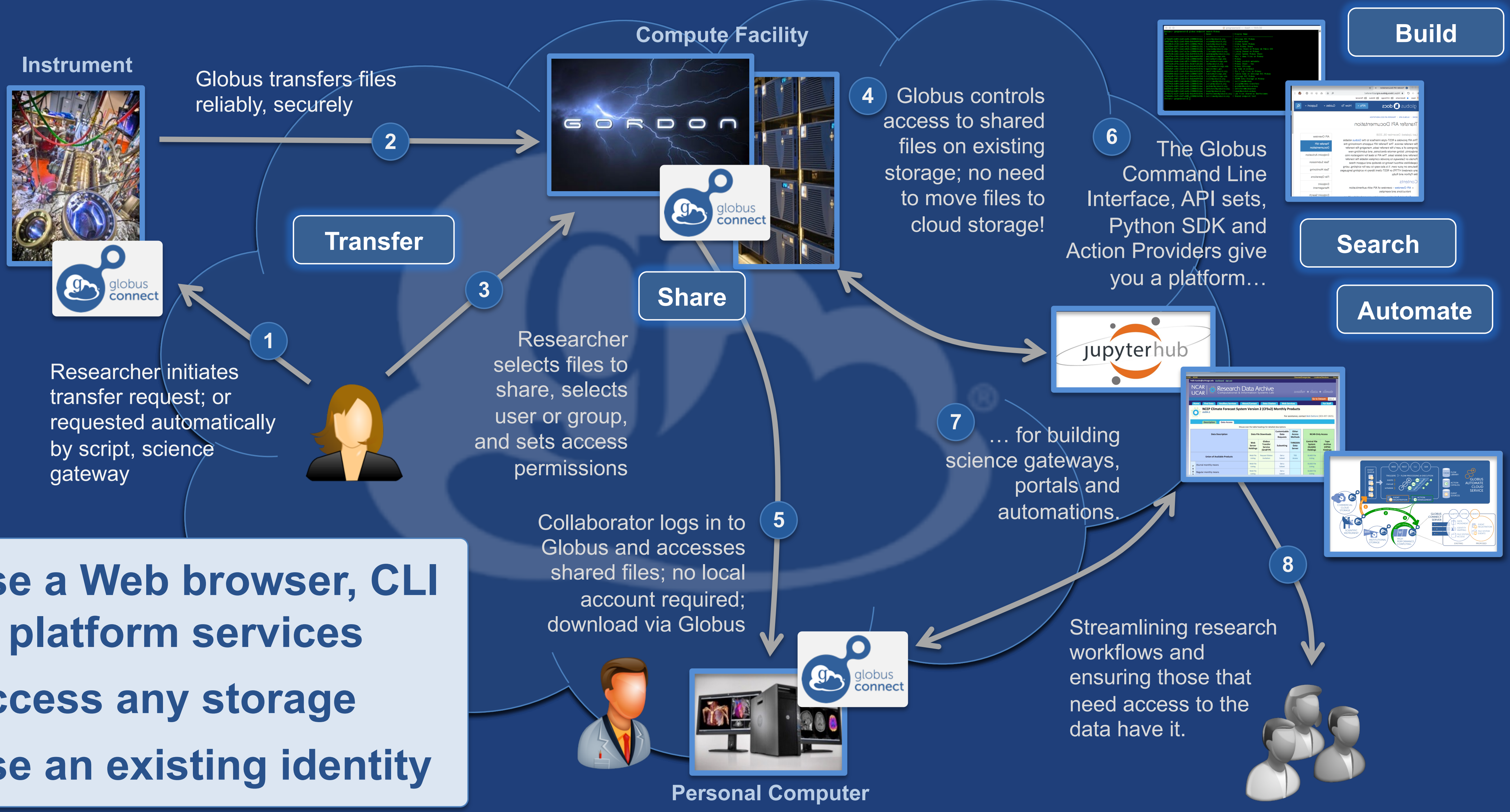
COMMERCIAL CLOUD STORAGE

RELIABLE TRANSFER 1PB PER DAY

1,600+ CONNECTED INSTITUTIONS

9,500+ ACTIVE SHARED ENDPOINTS

Globus and the research data lifecycle



Globus transfers files reliably, securely

1 Researcher initiates transfer request; or requested automatically by script, science gateway

2 Globus transfers files reliably, securely

3 Researcher selects files to share, selects user or group, and sets access permissions

4 Globus controls access to shared files on existing storage; no need to move files to cloud storage!

5 Collaborator logs in to Globus and accesses shared files; no local account required; download via Globus

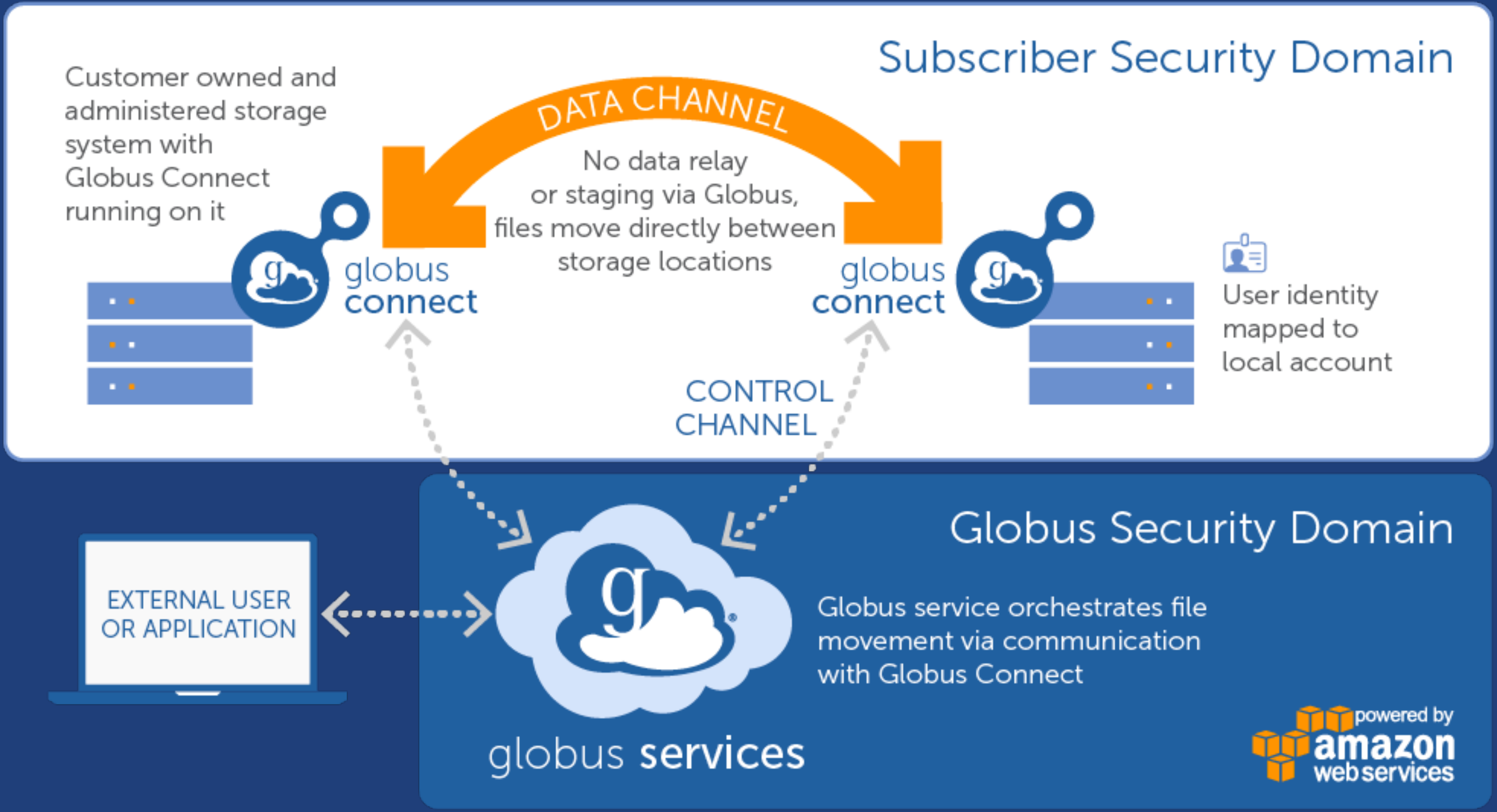
6 The Globus Command Line Interface, API sets, Python SDK and Action Providers give you a platform…

7 … for building science gateways, portals and automations.

8 Streamlining research workflows and ensuring those that need access to the data have it.

- Use a Web browser, CLI or platform services
- Access any storage
- Use an existing identity

Hybrid SaaS Architecture



Subscriber Security Domain
Customer owned and administered storage system with Globus Connect running on it
DATA CHANNEL
No data relay or staging via Globus, files move directly between storage locations
globus connect
globus connect
User identity mapped to local account
CONTROL CHANNEL
EXTERNAL USER OR APPLICATION
Globus Security Domain
Globus service orchestrates file movement via communication with Globus Connect
globus services
powered by amazon web services

Endpoints, Collections and Globus Connect

- **Globus Connect Server**
  - Multi-user Linux Systems

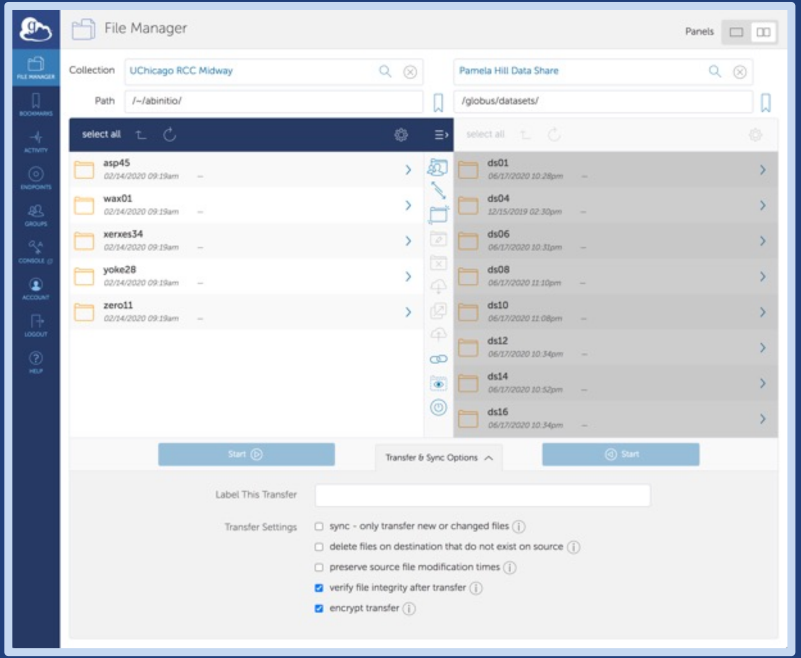
  docs.globus.org/globus-connect-server
- **Globus Connect Personal**
  - Personal workstations and laptops
  - OS specific instructions: docs.globus.org/how-to

The *ad hoc* user's perspective..

Use(r)-appropriate interfaces

Globus service

Web



File Manager

Collection: UChicago RCC Midway | Pamela Hill Data Share

Path: /~/abinitio/ | /globus/datasets/

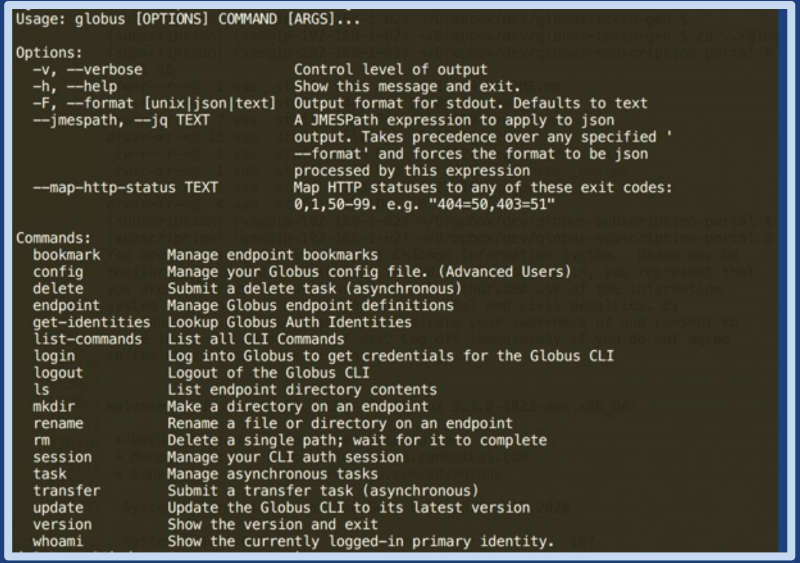
asp45 02/14/2020 09:19am –	ds01 06/17/2020 10:28pm –
wax01 02/14/2020 09:19am –	ds04 12/15/2019 02:30pm –
xerxes34 02/14/2020 09:19am –	ds06 06/17/2020 10:31pm –
yoke28 02/14/2020 09:19am –	ds08 06/17/2020 11:10pm –
zero11 02/14/2020 09:19am –	ds10 06/17/2020 11:08pm –
	ds12 06/17/2020 10:34pm –
	ds14 06/17/2020 10:52pm –
	ds16 06/17/2020 10:34pm –

Transfer & Sync Options

Label This Transfer

Transfer Settings
- □ sync - only transfer new or changed files
- □ delete files on destination that do not exist on source
- □ preserve source file modification times
- ☑ verify file integrity after transfer
- ☑ encrypt transfer

CLI



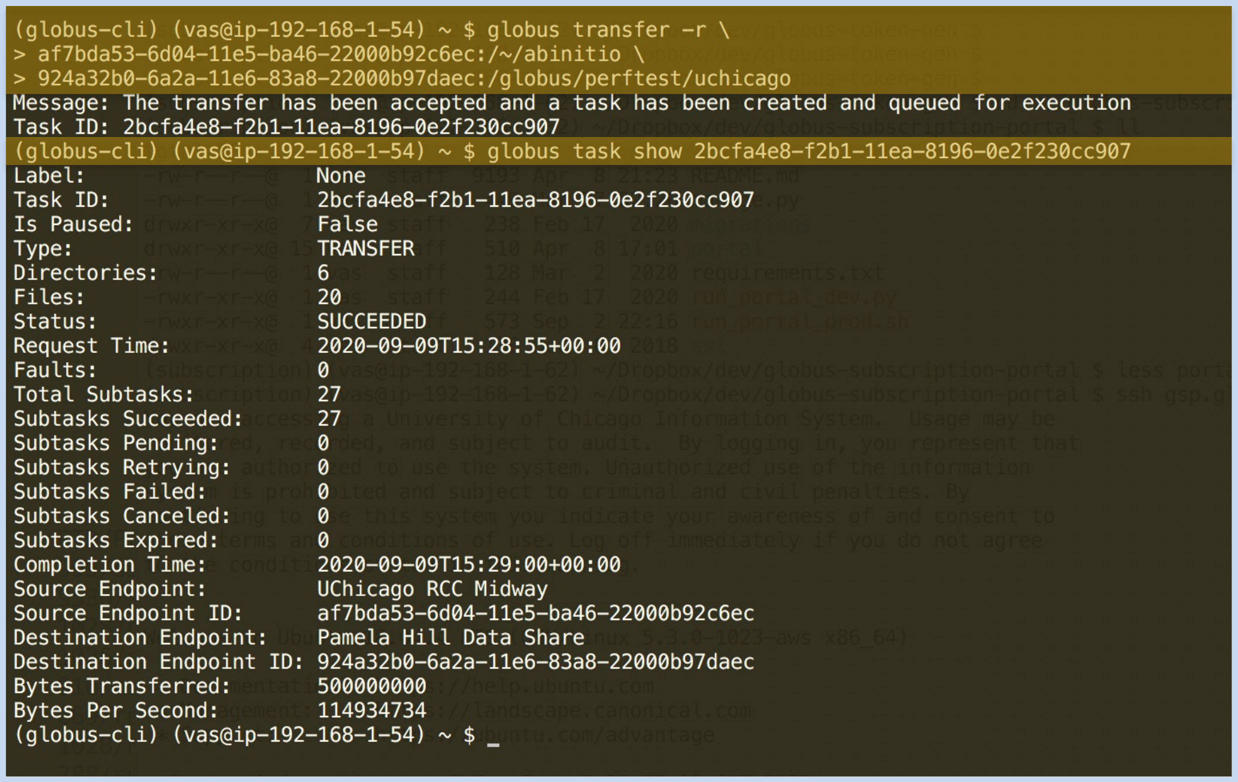
```
Usage: globus [OPTIONS] COMMAND [ARGS]...

Options:
  -v, --verbose                  Control level of output
  -h, --help                     Show this message and exit.
  -F, --format [unix|json|text]  Output format for stdout. Defaults to text
  --jmespath, --jq TEXT          A JMESPath expression to apply to json
                                 output. Takes precedence over any specified '
                                 --format' and forces the format to be json
                                 processed by this expression
  --map-http-status TEXT         Map HTTP statuses to any of these exit codes:
                                 0,1,50-99. e.g. "404=50,403=51"

Commands:
  bookmark       Manage endpoint bookmarks
  config         Manage your Globus config file. (Advanced Users)
  delete         Submit a delete task (asynchronous)
  endpoint       Manage Globus endpoint definitions
  get-identities Lookup Globus Auth Identities
  list-commands  List all CLI Commands
  login          Log into Globus to get credentials for the Globus CLI
  logout         Logout of the Globus CLI
  ls             List endpoint directory contents
  mkdir          Make a directory on an endpoint
  rename         Rename a file or directory on an endpoint
  rm             Delete a single path; wait for it to complete
  session        Manage your CLI auth session
  task           Manage asynchronous tasks
  transfer       Submit a transfer task (asynchronous)
  update         Update the Globus CLI to its latest version
  version        Show the version and exit
  whoami         Show the currently logged-in primary identity.
```

Platform (RESTful APIs)

```
GET /endpoint/go%23ep1
PUT /endpoint/demodoc#my_endpt
200 OK
X-Transfer-API-Version: 0.10
Content-Type: application/json
…
```

Globus CLI enables simple automation…

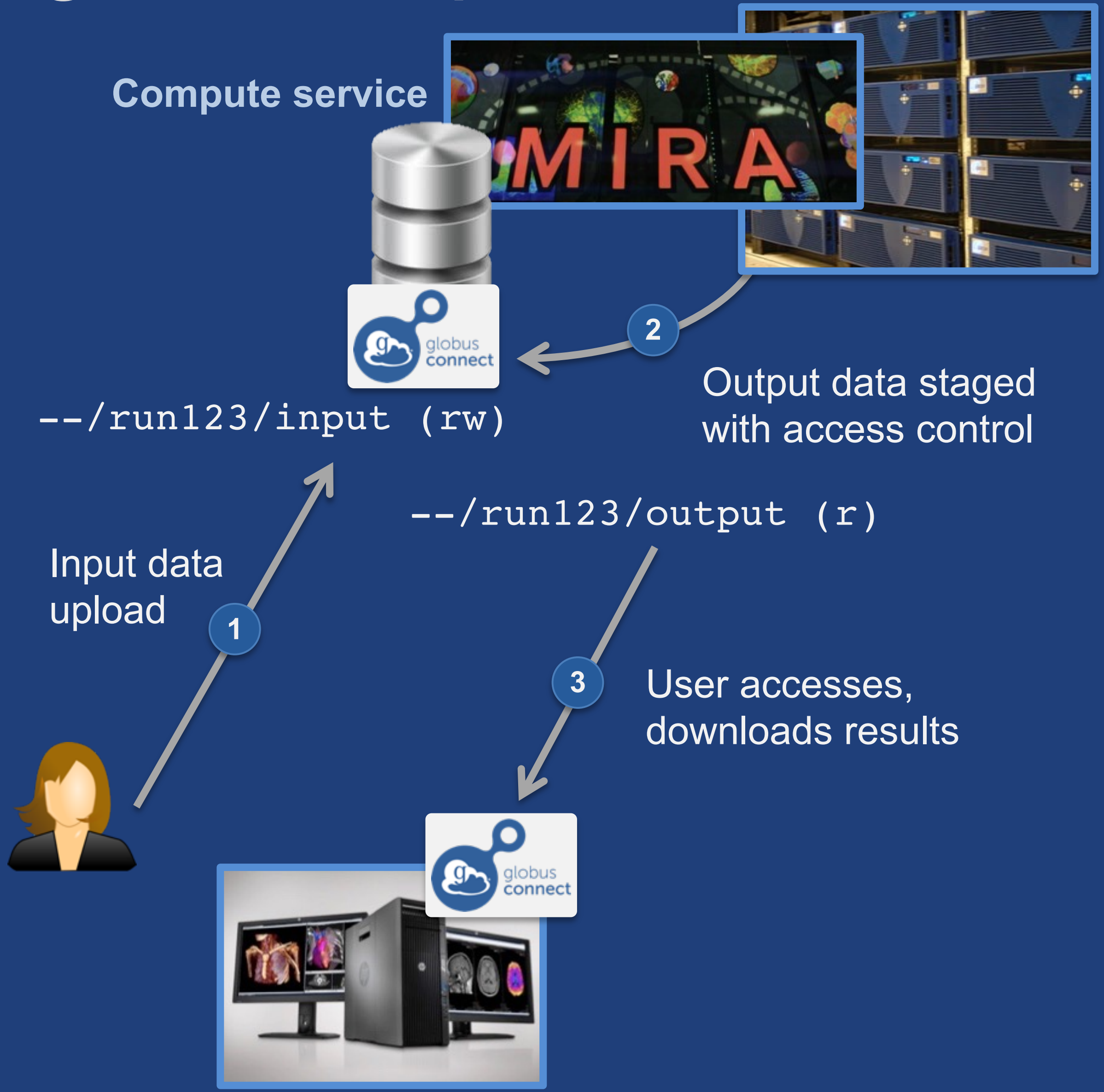
- **Open source native app**
- **Uses Python SDK**
- **Consistent access control model**
  - Access and refresh tokens
  - Tokens stored locally
- **docs.globus.org/cli**



```
(globus-cli) (vas@ip-192-168-1-54) ~ $ globus transfer -r \
> af7bda53-6d04-11e5-ba46-22000b92c6ec:/~/abinitio \
> 924a32b0-6a2a-11e6-83a8-22000b97daec:/globus/perftest/uchicago
Message: The transfer has been accepted and a task has been created and queued for execution
Task ID: 2bcfa4e8-f2b1-11ea-8196-0e2f230cc907
(globus-cli) (vas@ip-192-168-1-54) ~ $ globus task show 2bcfa4e8-f2b1-11ea-8196-0e2f230cc907
Label:                    None
Task ID:                  2bcfa4e8-f2b1-11ea-8196-0e2f230cc907
Is Paused:                False
Type:                     TRANSFER
Directories:              6
Files:                    20
Status:                   SUCCEEDED
Request Time:             2020-09-09T15:28:55+00:00
Faults:                   0
Total Subtasks:           27
Subtasks Succeeded:       27
Subtasks Pending:         0
Subtasks Retrying:        0
Subtasks Failed:          0
Subtasks Canceled:        0
Subtasks Expired:         0
Completion Time:          2020-09-09T15:29:00+00:00
Source Endpoint:          UChicago RCC Midway
Source Endpoint ID:       af7bda53-6d04-11e5-ba46-22000b92c6ec
Destination Endpoint:     Pamela Hill Data Share
Destination Endpoint ID:  924a32b0-6a2a-11e6-83a8-22000b97daec
Bytes Transferred:        500000000
Bytes Per Second:         114934734
(globus-cli) (vas@ip-192-168-1-54) ~ $ _
```

…such as data staging for compute

- User securely uploads data for analysis
- Results available with fine-grained permissions
- Automated setup/tear down of folders, permissions



BUT, the instruments are coming!

NEW YORK GENOME CENTER®

NISC

NIH Intramural Sequencing Center

CLOUDSAT

High Resolution

Cryo-EM

National Cryo-Electron Microscopy Facility

PURDUE

SCIENCE

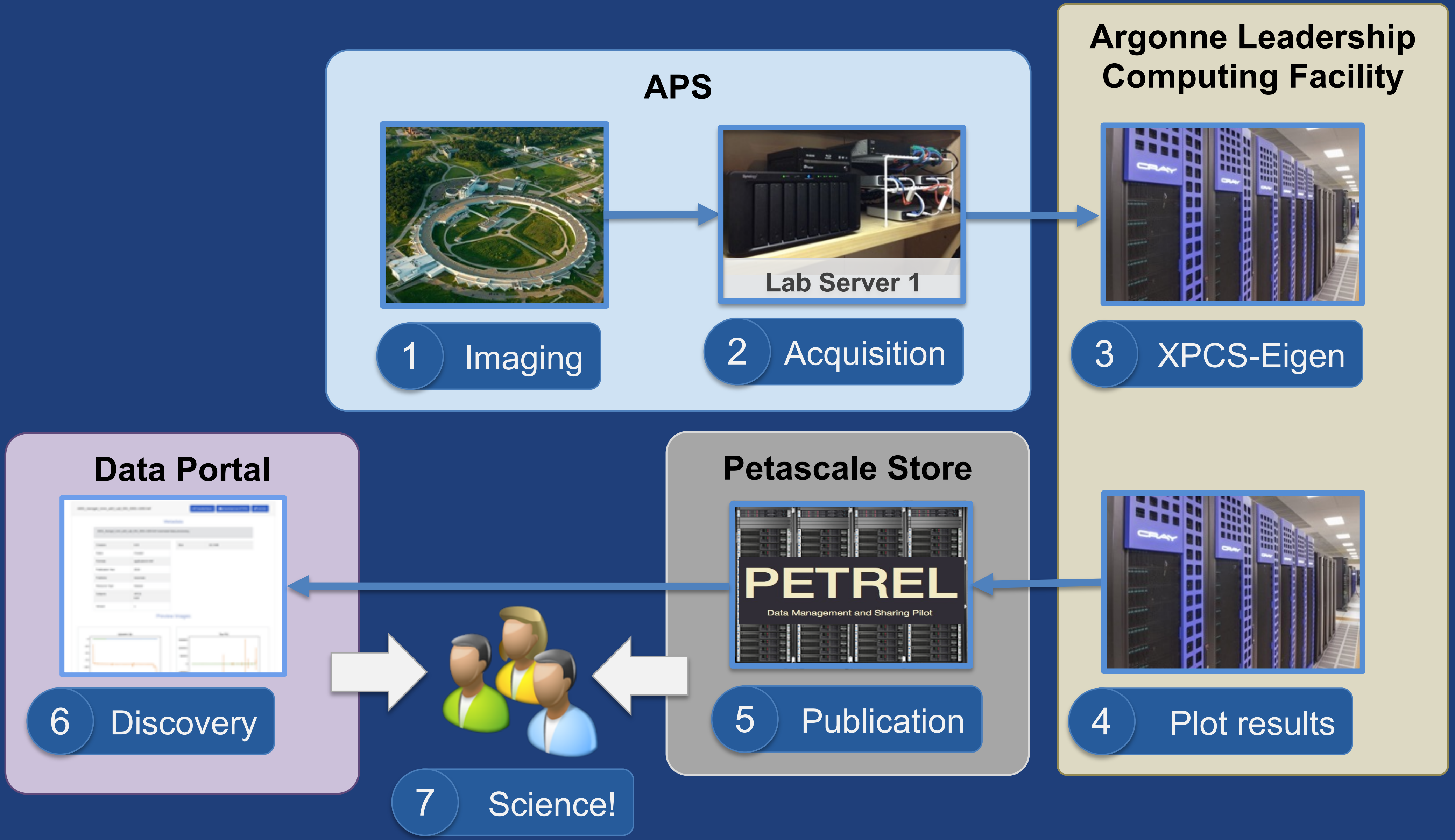
Purdue Cryo-EM Facility

ADVANCED LIGHT SOURCE

CLASS

Image source: GOES-16 and S-NPP

Automating XPCS at scale



APS

Lab Server 1

1 Imaging

2 Acquisition

Argonne Leadership Computing Facility

3 XPCS-Eigen

4 Plot results

Petascale Store

PETREL

Data Management and Sharing Pilot

5 Publication

Data Portal

6 Discovery

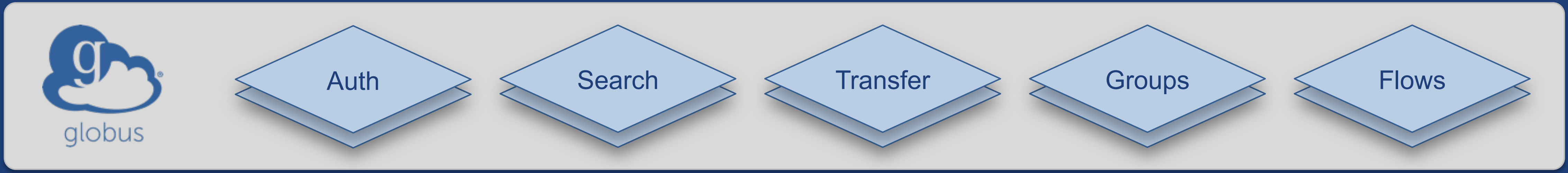
7 Science!

Making data discoverable

petreldata.alcf.anl.gov

Instrument data orchestration

- Authentication and Authorization
- Data transfer and sharing
- Data description and discovery
- Data (and compute) orchestration



globus

Auth

Search

Transfer

Groups

Flows

Globus Auth: Foundational IAM service

Auth

Brokers authentication and authorization among…

- End-users
- Identity providers: enterprise, external (federated identities)
- Services: resource servers with REST APIs
- Apps: web, mobile, desktop, command line clients
- Services acting as clients to other services

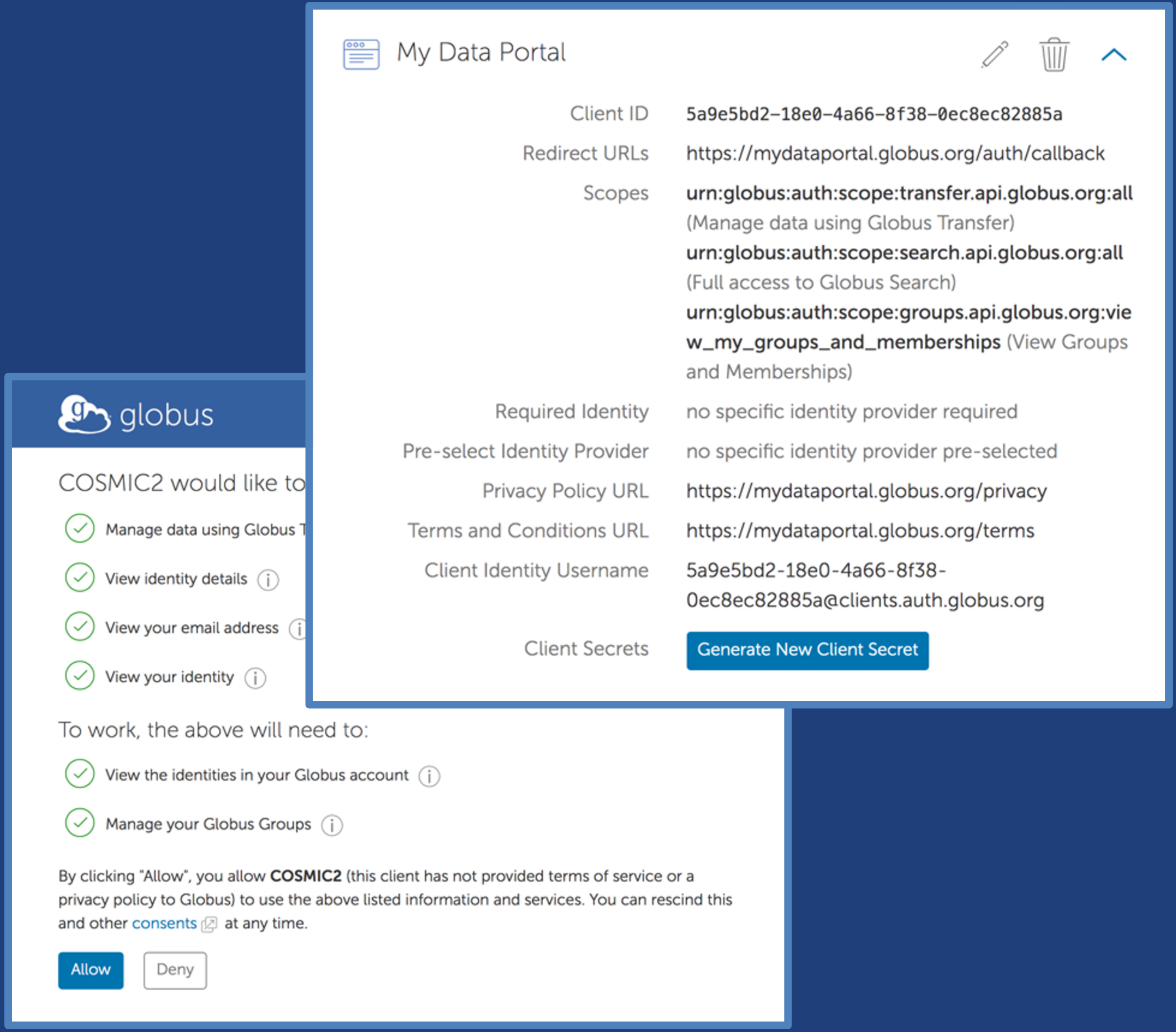
- **OAuth 2.0 Authorization Framework (a.k.a. OAuth2)**
- **OpenID Connect Core 1.0 (a.k.a. OIDC)**

Step 0: Application registration

Auth

- **Set desired scopes**
- **Set callback URL**
- **Get client ID and secret**
- **Consents implement least privileges principle**

developers.globus.org



My Data Portal

Client ID	5a9e5bd2-18e0-4a66-8f38-0ec8ec82885a
Redirect URLs	https://mydataportal.globus.org/auth/callback
Scopes	**urn:globus:auth:scope:transfer.api.globus.org:all** (Manage data using Globus Transfer) **urn:globus:auth:scope:search.api.globus.org:all** (Full access to Globus Search) **urn:globus:auth:scope:groups.api.globus.org:view_my_groups_and_memberships** (View Groups and Memberships)
Required Identity	no specific identity provider required
Pre-select Identity Provider	no specific identity provider pre-selected
Privacy Policy URL	https://mydataportal.globus.org/privacy
Terms and Conditions URL	https://mydataportal.globus.org/terms
Client Identity Username	5a9e5bd2-18e0-4a66-8f38-0ec8ec82885a@clients.auth.globus.org
Client Secrets	Generate New Client Secret

globus

COSMIC2 would like to

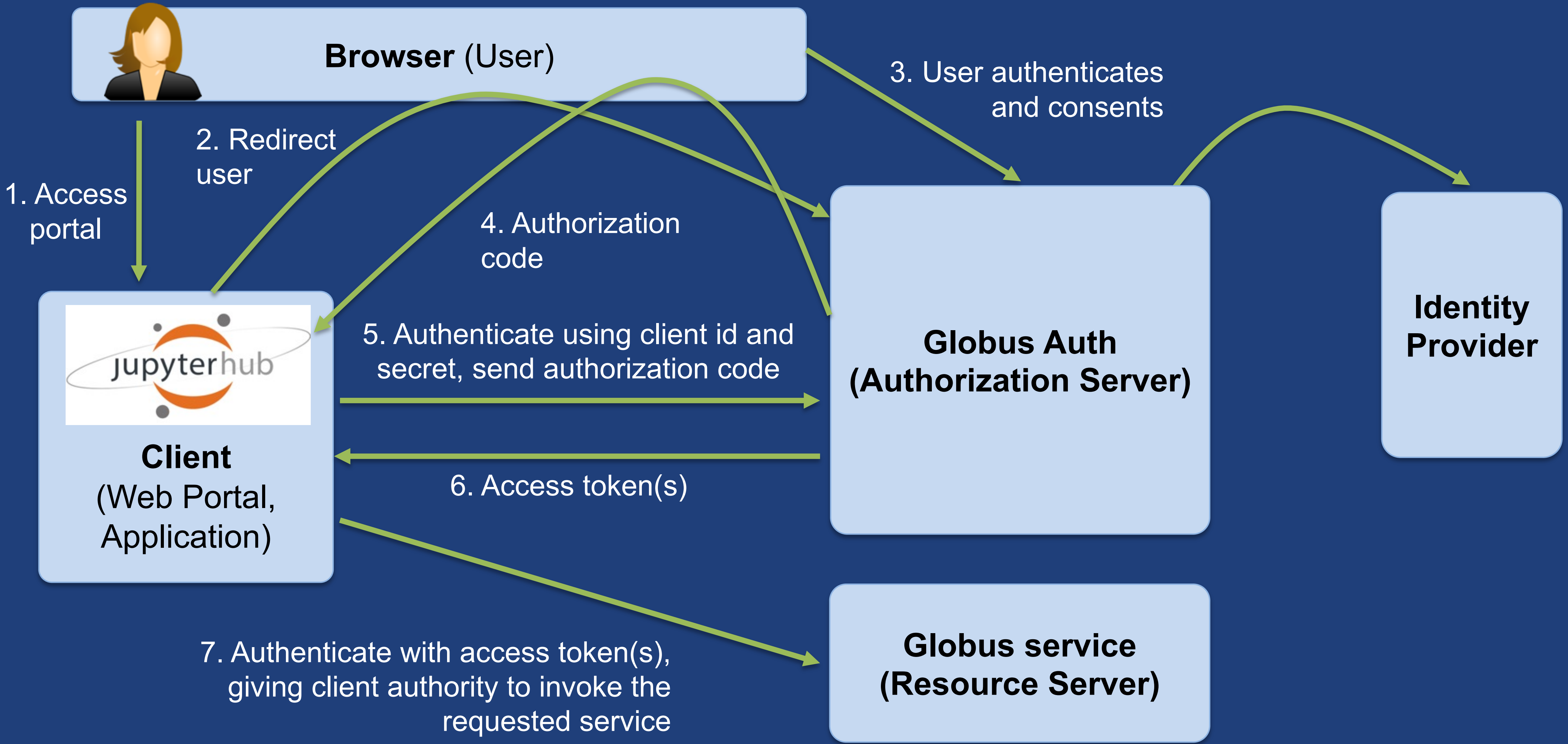
- ✓ Manage data using Globus T
- ✓ View identity details
- ✓ View your email address
- ✓ View your identity

To work, the above will need to:

- ✓ View the identities in your Globus account
- ✓ Manage your Globus Groups

By clicking "Allow", you allow **COSMIC2** (this client has not provided terms of service or a privacy policy to Globus) to use the above listed information and services. You can rescind this and other consents at any time.

Authorization Code Grant



Browser (User)

1. Access portal

2. Redirect user

3. User authenticates and consents

4. Authorization code

5. Authenticate using client id and secret, send authorization code

6. Access token(s)

7. Authenticate with access token(s), giving client authority to invoke the requested service

Globus Python SDK

- **Python client library for accessing Globus APIs**
- **For example**, `globus_sdk.TransferClient` **class handles connection management, security, framing, marshaling when accessing the Transfer service**

```
from globus_sdk import TransferClient, AuthClient
tc = TransferClient()
```

globus.github.io/globus-sdk-python

Experimenting with Globus platform services

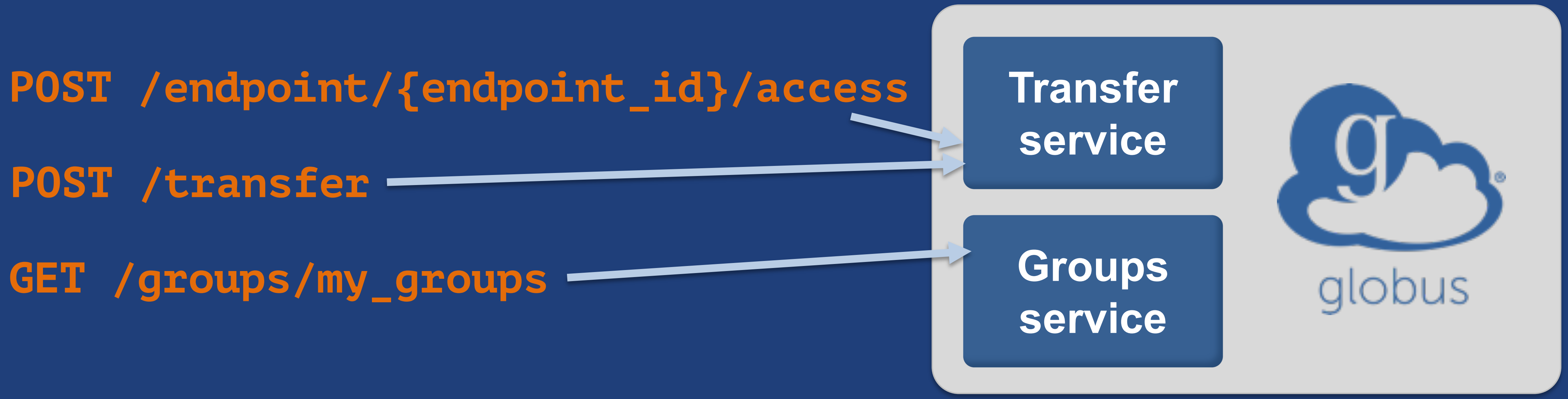
jupyter.demo.globus.org

Data transfer and sharing

Transfer

Groups

- Move data to collection → Submit Transfer task
- Make data accessible → Set guest collection access rule
- Grant user(s) access → Add/confirm Group membership



```
POST /endpoint/{endpoint_id}/access

POST /transfer

GET /groups/my_groups
```

Data description and discovery

Search

- Metadata store with fine-grained visibility controls
- Schema agnostic
  → dynamic schemas
- Simple search using URL query parameters
- Complex search using search request document

docs.globus.org/api/search

Data ingest with Globus Search

POST /index/{index_id}/ingest'

```
{
  "ingest_type": "GMetaList",
  "ingest_data": {
  "gmeta": [
    {
      "id": "filetype",
      "subject”: "https://search.api.globus.org/abc.txt",
      "visible_to": ["public"],
      "content": {
        "metadata-schema/file#type": "file”
      }
    },
    ...
  ]
}
```

Search Index
globus

Data ingest with Globus Search

POST /index/{index_id}/ingest'

```
{
  "ingest_type": "GMetaList",
  "ingest_data": {
  "gmeta": [
    {
      "id": "size",
      "subject": "https://search.api.globus.org/abc.txt",
      "visible_to": ["urn:globus:auth:identity:46bd0f56-
                        e24f-11e5-a510-131bef46955c"],
      "content": {
        "metadata-schema/file#size": "1000000",
        "metadata-schema/file#size_human": "1MB”
      }
    },
    ...
  ]
}
```

Search Index
globus

Visibility limited to Globus Auth identity
- Single user
- Globus Group
- Registered client application

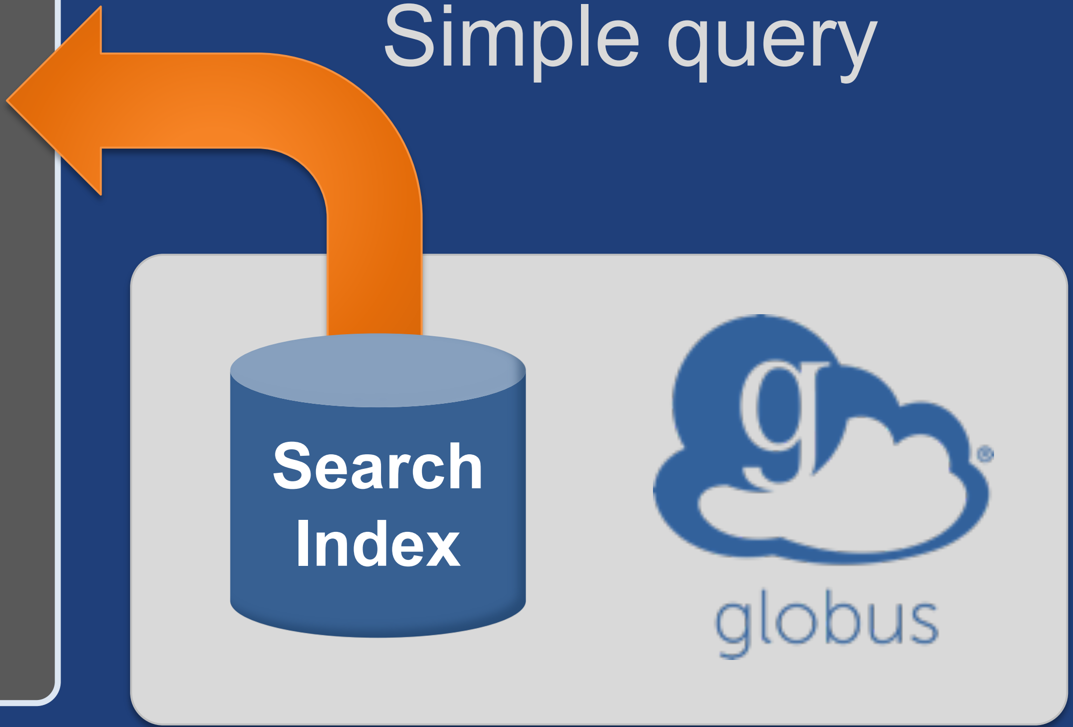
Data discovery with Globus Search

Search

GET /index/{index_id}/search?q=type%3Ahdf5

```
{
  "@datatype": "GSearchResult",
  "@version": "2017-09-01",
  "count": 1,
  "gmeta": [
    {
      "@datatype": "GMetaResult",
      "@version": "2019-08-27",
      "entries": [
        { ... }
      ],
      "subject": "https://..."
    }
  ],
  "offset": 0,
  "total": 1
}
```

Simple query



Data discovery with Globus Search

POST /index/{index_id}/search

```
{
  "filters": [
    {
      "type": "range",
      "field_name": ”pubdate",
      "values": [
        {
          "from": "*",
          "to": "2020-12-31"
        }
      ]
    }
  ],
  "facets": [
    {
      "name": "Publication Date",
      "field_name": "pubdate",
      ...
    }
  ]
}
```

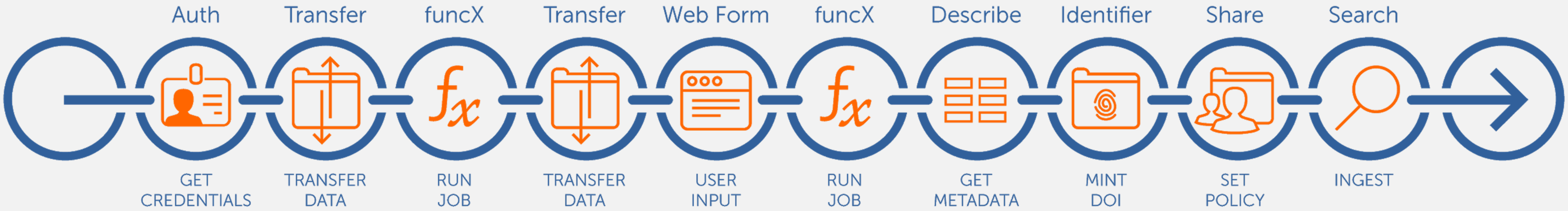
Complex query

Search Index
globus

Data (and compute) automation

- **Flows**: A platform service for defining, applying, and sharing distributed research automation flows
- Flows comprise **Actions**
- **Action Providers**: Called by Flows to perform tasks
- **Triggers***: Start flows based on events

* In development



globus.org

docs.globus.org

outreach@globus.org

support@globus.org